

# UVA CS 4501 - 001 / 6501 – 007

## Introduction to Machine Learning and Data Mining

### Lecture 9: Classification with Support Vector Machine (cont.)

Yanjun Qi / Jane

University of Virginia  
Department of  
Computer Science

Where we are ? →

Five major sections of this course

- ~~Regression (supervised)~~
- Classification (supervised)
- Unsupervised models
- Learning theory
- Graphical models

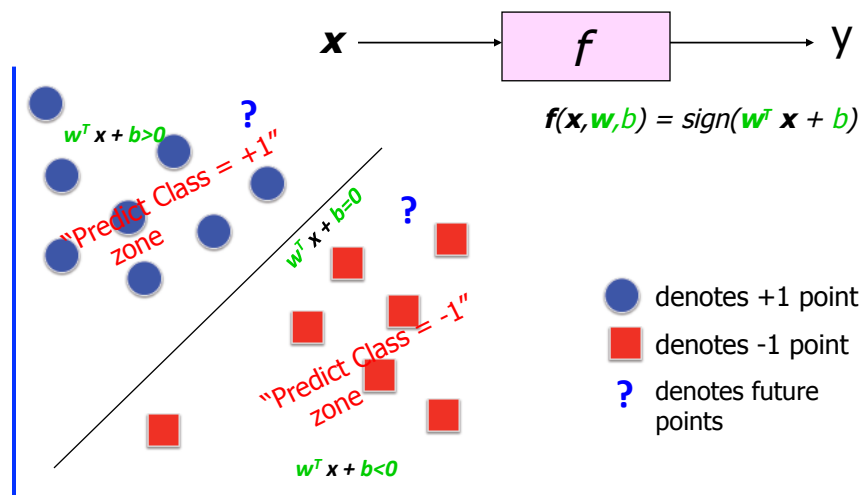
## Today

### □ Review of Classification

#### □ Support Vector Machine (SVM)

- ✓ Large Margin Linear Classifier
- ✓ Define Margin (M) in terms of model parameter
- ✓ Optimization to learn model parameters (w, b)
- ✓ Non linearly separable case
- ✓ Optimization with dual form

### e.g. SUPERVISED Linear Binary Classifier



## Types of classifiers

- We can divide the large variety of classification approaches into **roughly three major types**

### 1. Discriminative

- directly estimate a decision rule/boundary
- e.g., **support vector machine**, decision tree

### 2. Generative:

- build a generative statistical model
- e.g., Bayesian networks

### 3. Instance based classifiers

- Use observation directly (no models)
- e.g. K nearest neighbors

## A study comparing Classifiers

→ 11 binary classification problems / 8 metrics

Table 2. Normalized scores for each learning algorithm by metric (average over eleven problems)

MODEL	CAL	ACC	FSC	LFT	ROC	APR	BEP	RMS	MXE	MEAN	OPT-SEL
BST-DT	PLT	.843*	.779	<b>.939</b>	<b>.963</b>	<b>.938</b>	.929*	<b>.880</b>	<b>.896</b>	<b>.896</b>	<b>.917</b>
RF	PLT	.872*	.805	.934*	.957	.931	<b>.930</b>	.851	.858	.892	.898
BAG-DT	-	.846	.781	.938*	.962*	.937*	.918	.845	.872	<b>.887*</b>	.899
BST-DT	ISO	.826*	.860*	.929*	.952	.921	.925*	.854	.815	.885	.917*
RF	-	<b>.872</b>	.790	.934*	.957	.931	<b>.930</b>	.829	.830	.884	.890
BAG-DT	PLT	.841	.774	.938*	.962*	.937*	.918	.836	.852	.882	.895
RF	ISO	.861*	<b>.861</b>	.923	.946	.910	.925	.836	.776	.880	.895
BAG-DT	ISO	.826	<b>.843*</b>	<b>.933*</b>	.954	.921	.915	.832	.791	.877	.894
SVM	PLT	.824	.760	.895	.938	.898	.913	.831	.836	.862	.880
ANN	-	.803	.762	.910	.936	.892	.899	.811	.821	.854	.885
SVM	ISO	.813	<b>.836*</b>	.892	.925	.882	.911	.814	.744	.852	.882
ANN	PLT	.815	.748	.910	.936	.892	.899	.783	.785	.846	.875
ANN	ISO	.803	.836	.908	.924	.876	.891	.777	.718	.842	.884
BST-DT	-	<b>.834*</b>	.816	<b>.939</b>	<b>.963</b>	<b>.938</b>	.929*	.598	.605	.828	.851
KNN	PLT	.757	.707	.889	.918	.872	.872	.742	.764	.815	.837
KNN	-	.756	.728	.889	.918	.872	.872	.729	.718	.810	.830
KNN	ISO	.755	.758	.882	.907	.854	.869	.738	.706	.809	.844
BST-STMP	PLT	.724	.651	.876	.908	.853	.845	.716	.754	.791	.808
SVM	-	.817	.804	.895	.938	.899	.913	.514	.467	.781	.810
BST-STMP	ISO	.709	.744	.873	.899	.835	.840	.695	.646	.780	.810
BST-STMP	-	.741	.684	.876	.908	.853	.845	.394	.382	.710	.726
DT	ISO	.648	.654	.818	.838	.756	.778	.590	.589	.709	.774

Yanjun Qi / UVA CS 4501-01-6501-07

$X_1$	$X_2$	$X_3$	$Y$

## A Dataset for **binary** classification

$$f : X \rightarrow Y$$

$X \rightarrow Y$

Output as  
Binary Class:  
only two  
possibilities

- **Data/points/instances/examples/samples/records:** [ rows ]
- **Features/attributes/dimensions/independent variables/covariates/predictors/regressors:** [ columns, except the last ]
- **Target/outcome/response/label/dependent variable:** special column to be predicted [ last column ]

9/25/14
7

Yanjun Qi / UVA CS 4501-01-6501-07

## History of SVM

- SVM is inspired from statistical learning theory [3]
- SVM was first introduced in 1992 [1]
- SVM becomes popular because of its success in handwritten digit recognition
  - 1.1% test error rate for SVM. This is the same as the error rates of a carefully constructed neural network, LeNet 4.
    - See Section 5.11 in [2] or the discussion in [3] for details
- SVM is now regarded as an important example of “kernel methods”, arguably the hottest area in machine learning ten years ago

[1] B.E. Boser *et al.* A Training Algorithm for Optimal Margin Classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory 5 144-152, Pittsburgh, 1992.  
 [2] L. Bottou *et al.* Comparison of classifier methods: a case study in handwritten digit recognition. Proceedings of the 12th IAPR International Conference on Pattern Recognition, vol. 2, pp. 77-82, 1994.  
 [3] V. Vapnik. The Nature of Statistical Learning Theory. 2<sup>nd</sup> edition, Springer, 1999.

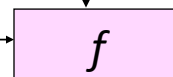
9/25/14
8

## Today

- ☐ Review of Classification
- ☐ **Support Vector Machine (SVM)**
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ✓ Optimization with dual form

## Linear Classifiers

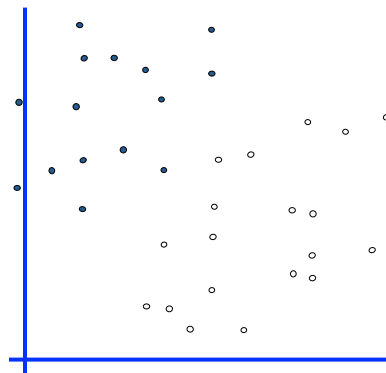
$\mathbf{x}$



$y^{\text{est}}$

$$f(\mathbf{x}, \mathbf{w}, b) = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$$

- denotes +1
- denotes -1



How would you classify this data?

Yanjun Qi / UVA CS 4501-01-6501-07

# Linear Classifiers

$x$  →  $f$  →  $y^{\text{est}}$

$f(x, w, b) = \text{sign}(w^T x + b)$

• denotes +1  
○ denotes -1

How would you classify this data?

9/25/14 11

Yanjun Qi / UVA CS 4501-01-6501-07

# Linear Classifiers

$x$  →  $f$  →  $y^{\text{est}}$

$f(x, w, b) = \text{sign}(w^T x + b)$

• denotes +1  
○ denotes -1

How would you classify this data?

9/25/14 12

Yanjun Qi / UVA CS 4501-01-6501-07

## Linear Classifiers

$x$

$f$

$y^{\text{est}}$

$f(x, w, b) = \text{sign}(w^T x + b)$

- denotes +1
- denotes -1

How would you classify this data?

9/25/14
13

Yanjun Qi / UVA CS 4501-01-6501-07

## Linear Classifiers

$x$

$f$

$y^{\text{est}}$

$f(x, w, b) = \text{sign}(w^T x + b)$

- denotes +1
- denotes -1

Any of these would be fine..

..but which is best?

9/25/14
14

Yanjun Qi / UVA CS 4501-01-6501-07

## Classifier Margin

$x$

$y^{\text{est}}$

$$f(x, w, b) = \text{sign}(w^T x + b)$$

- denotes +1
- denotes -1

Define the **margin** of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

9/25/14
15

Yanjun Qi / UVA CS 4501-01-6501-07

## Maximum Margin

$x$

$y^{\text{est}}$

$$f(x, w, b) = \text{sign}(w^T x + b)$$

- denotes +1
- denotes -1

The **maximum margin linear classifier** is the linear classifier with the, um, maximum margin. This is the simplest kind of SVM (Called an LSVM)

9/25/14
16



Yanjun Qi / UVA CS 4501-01-6501-07

## Maximum Margin

$x$

• denotes +1  
○ denotes -1

Support Vectors are those datapoints that the margin pushes up against

$\alpha$

$f$

$y_{est}$

$f(x, w, b) = \text{sign}(w^T x + b)$

The **maximum margin linear classifier** is the linear classifier with the, maximum margin. This is the simplest kind of SVM (Called an LSVM)

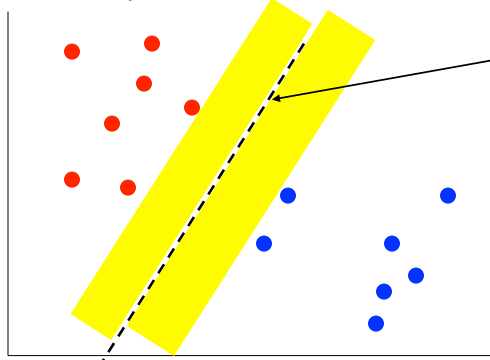
Linear SVM

9/25/14
17

Yanjun Qi / UVA CS 4501-01-6501-07

## Max margin classifiers

- Instead of fitting all points, focus on boundary points
- Learn a boundary that leads to the largest margin from both sets of points

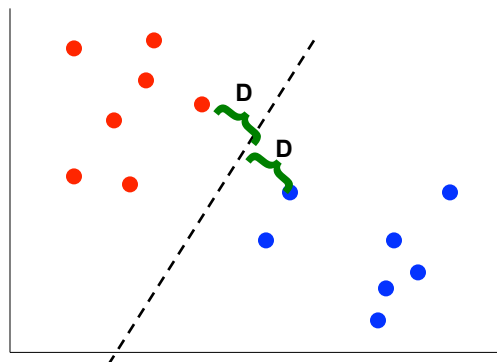


From all the possible boundary lines, this leads to the largest margin on both sides

9/25/14
18

## Max margin classifiers

- Instead of fitting all points, focus on boundary points
- Learn a boundary that leads to the largest margin from points on both sides



Why?

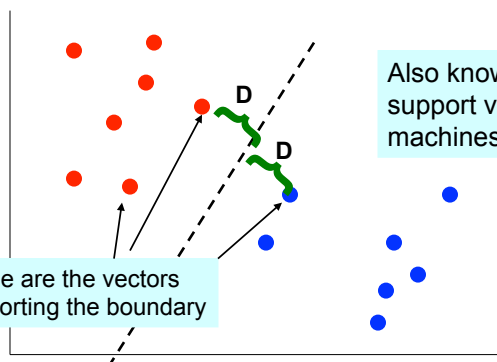
- Intuitive, 'makes sense'
- Some theoretical support
- Works well in practice

9/25/14

19

## Max margin classifiers

- Instead of fitting all points, focus on boundary points
- Learn a boundary that leads to the largest margin from points on both sides



Also known as linear support vector machines (SVMs)

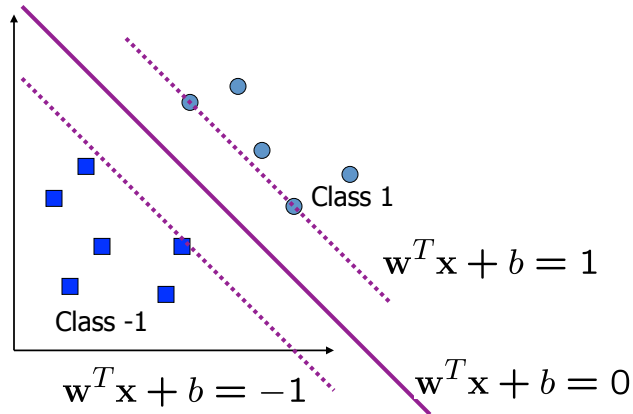
These are the vectors supporting the boundary

9/25/14

20

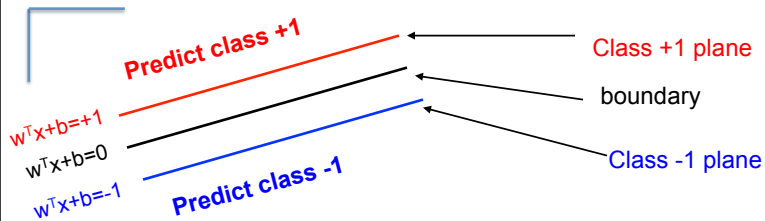
## Max-margin & Decision Boundary

- The decision boundary should be as far away from the data of both classes as possible



## Specifying a max margin classifier

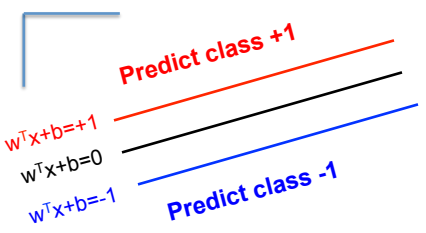
Yanjun Qi / UVA CS 4501-01-6501-07



Classify as +1	if	$w^T x + b \geq 1$
Classify as -1	if	$w^T x + b \leq -1$
Undefined	if	$-1 < w^T x + b < 1$

YanJun Qi / UVA CS 4501-01-6501-07

## Specifying a max margin classifier



Is the linear separation assumption realistic?


We will deal with this shortly, but let's assume it for now

Classify as +1	if	$w^T x + b \geq 1$
Classify as -1	if	$w^T x + b \leq -1$
Undefined	if	$-1 < w^T x + b < 1$

9/25/1423

YanJun Qi / UVA CS 4501-01-6501-07

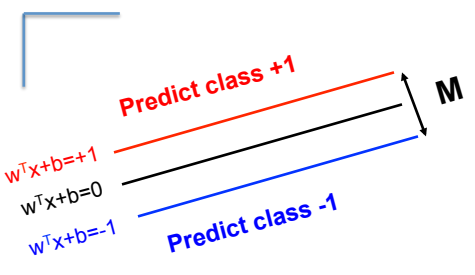
## Today

- Review of Classification
- Support Vector Machine (SVM)**
  - ✓ Large Margin Linear Classifier
  -  ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ✓ Optimization with dual form

9/25/1424

Yanjun Qi / UVA CS 4501-01-6501-07

## Maximizing the margin



Classify as +1 if  $w^T x + b \geq 1$

Classify as -1 if  $w^T x + b \leq -1$

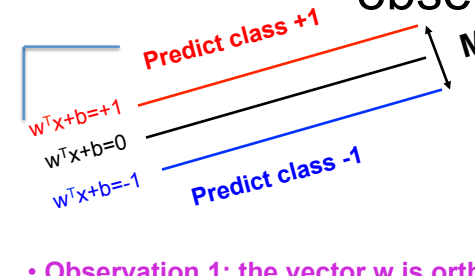
Undefined if  $-1 < w^T x + b < 1$

- Lets define the width of the margin by  $M$
- How can we encode our goal of maximizing  $M$  in terms of our parameters ( $w$  and  $b$ )?
- Lets start with a few observations

9/25/1425

Yanjun Qi / UVA CS 4501-01-6501-07

## Maximizing the margin: observation-1



Classify as +1 if  $w^T x + b \geq 1$

Classify as -1 if  $w^T x + b \leq -1$

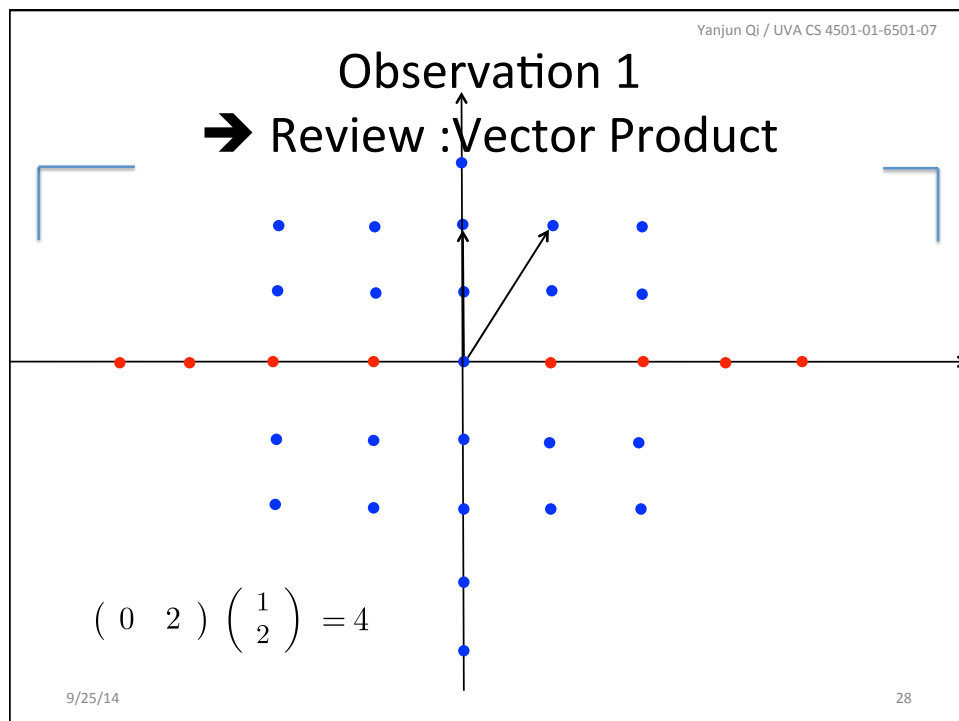
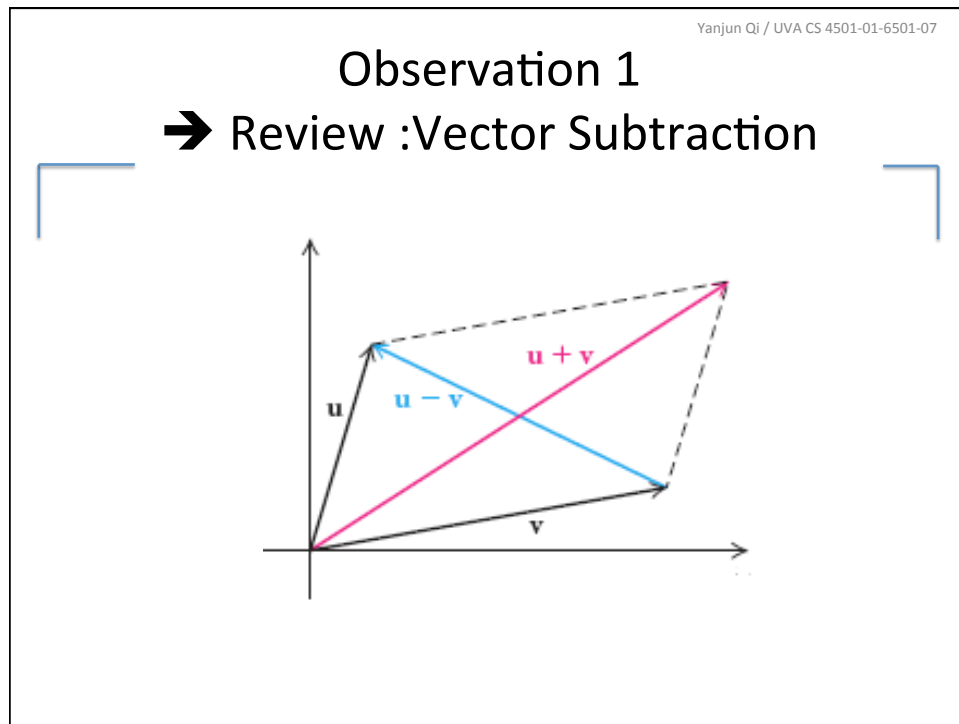
Undefined if  $-1 < w^T x + b < 1$

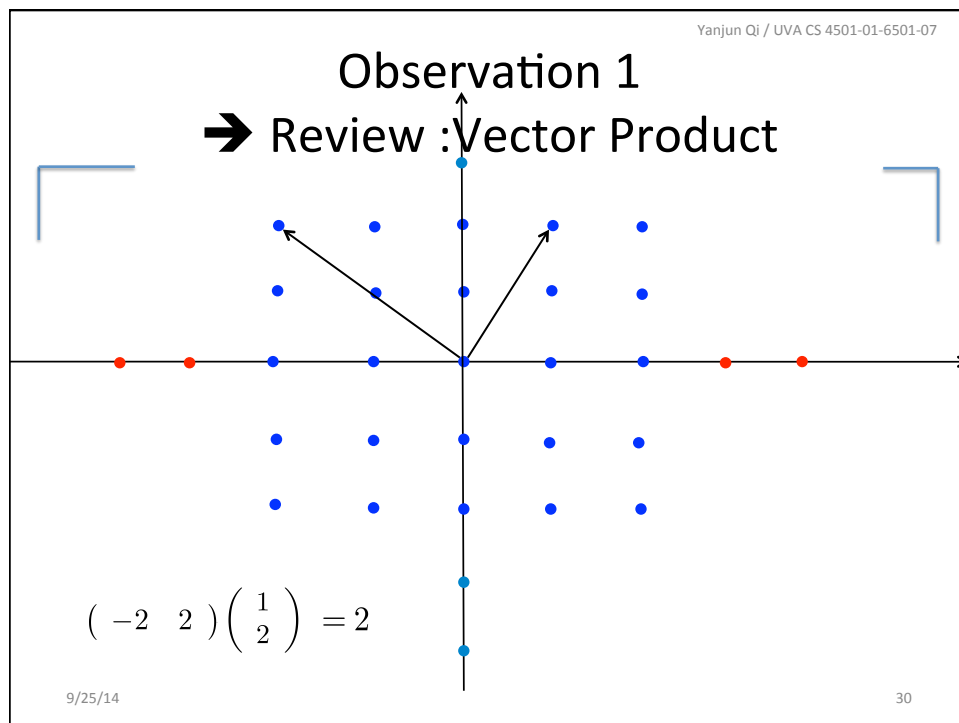
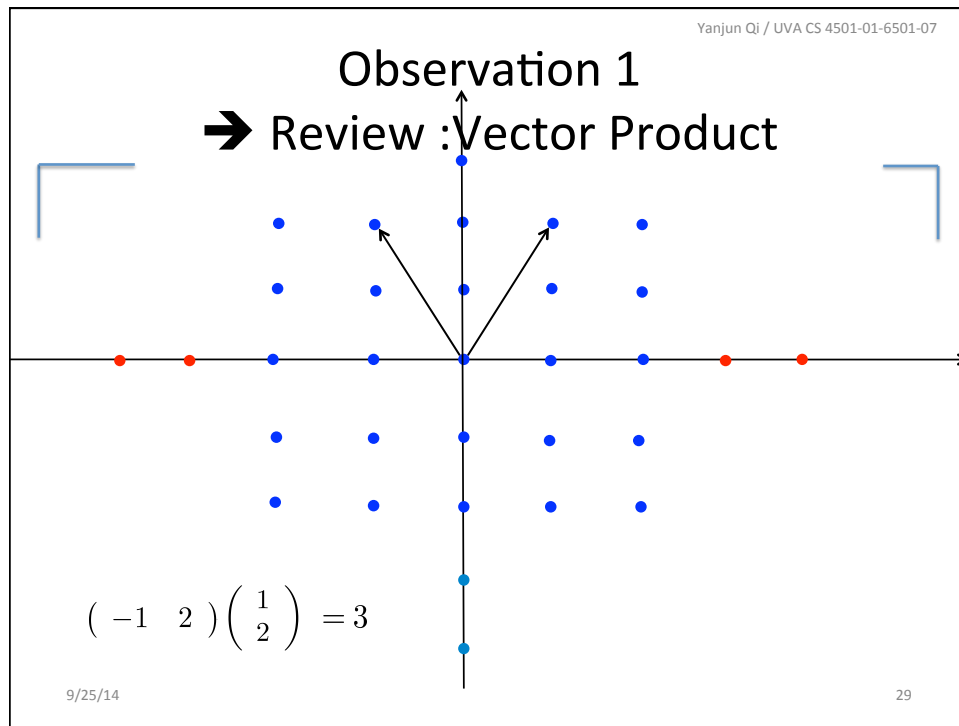
- **Observation 1: the vector  $w$  is orthogonal to the +1 plane**
- Why?

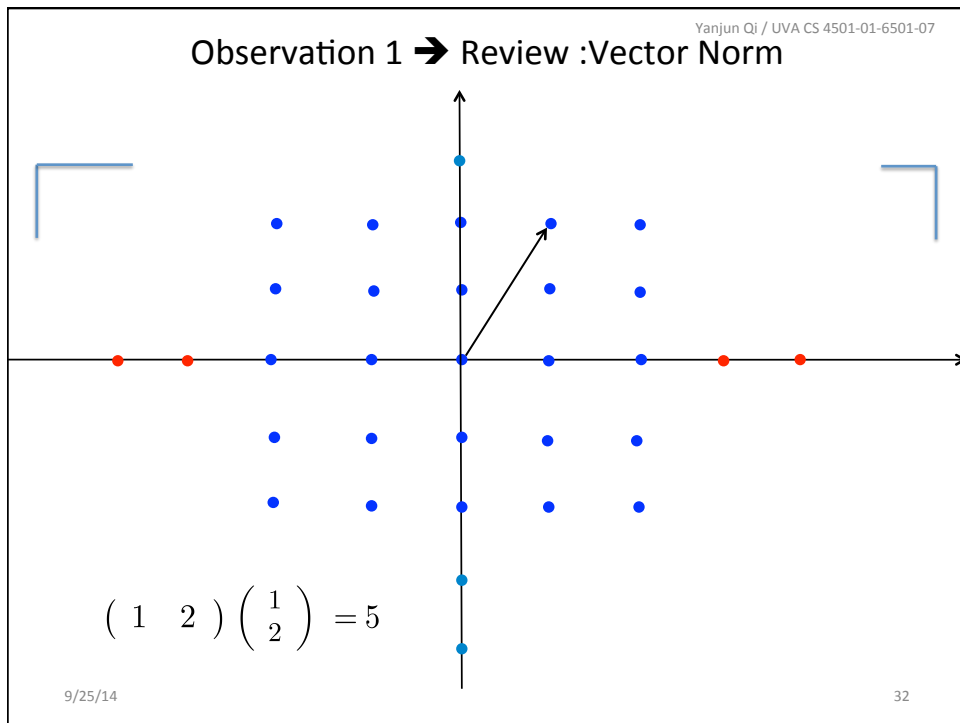
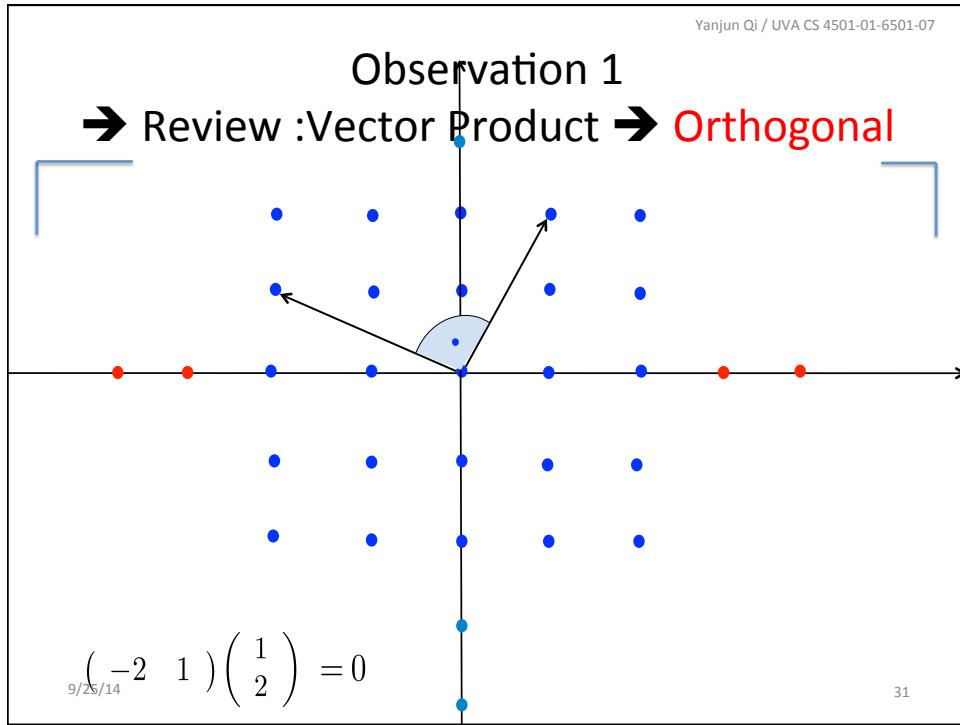
Let  $u$  and  $v$  be two points on the +1 plane, then for the vector defined by  $u$  and  $v$  we have  $w^T(u-v) = 0$

Corollary: the vector  $w$  is orthogonal to the -1 plane

9/25/1426









## Observation 1 → Review : Vector Product, Orthogonal, and Norm

For two vectors  $x$  and  $y$ ,  

$$x^T y$$
 is called the (*inner*) *vector product*.

$x$  and  $y$  are called *orthogonal* if  

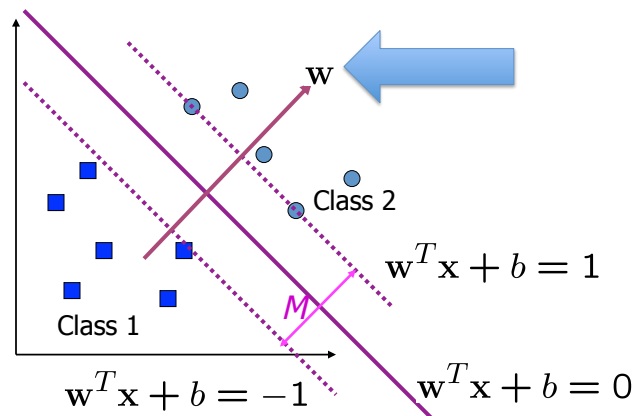
$$x^T y = 0$$

The square root of the product of a vector with itself,  

$$\sqrt{x^T x}$$
 is called the *2-norm* ( $\|x\|_2$ ), can also write as  $|x|$

## Maximizing the margin: observation-1

- Observation 1: the vector  $w$  is orthogonal to the +1 plane



Yanjun Qi / UVA CS 4501-01-6501-07

## Maximizing the margin: observation-1

Classify as +1 if  $w^T x + b \geq 1$

Classify as -1 if  $w^T x + b \leq -1$

Undefined if  $-1 < w^T x + b < 1$

- Observation 1: the vector  $w$  is orthogonal to the +1 plane
- Why?

Let  $u$  and  $v$  be two points on the +1 plane,  
then for the vector defined by  $u$  and  $v$  we have  
 $w^T(u-v) = 0$

**Corollary: the vector  $w$  is orthogonal to the -1 plane**

9/25/14
35

Yanjun Qi / UVA CS 4501-01-6501-07

## Maximizing the margin: observation-2

Classify as +1 if  $w^T x + b \geq 1$

Classify as -1 if  $w^T x + b \leq -1$

Undefined if  $-1 < w^T x + b < 1$

- Observation 1: the vector  $w$  is orthogonal to the +1 and -1 planes
- Observation 2: if  $x^+$  is a point on the +1 plane and  $x^-$  is the closest point to  $x^+$  on the -1 plane then

$$x^+ = \lambda w + x^-$$

Since  $w$  is orthogonal to both planes  
we need to 'travel' some distance  
along  $w$  to get from  $x^+$  to  $x^-$

9/25/14
36

Yanjun Qi / UVA CS 4501-01-6501-07

## Putting it together

- $w^T x^+ + b = +1$
- $w^T x^- + b = -1$
- $x^+ = \lambda w + x^-$
- $|x^+ - x^-| = M$

We can now define M in terms of w and b

$$w^T x^+ + b = +1$$

$$\Rightarrow w^T (\lambda w + x^-) + b = +1$$

$$\Rightarrow w^T x^- + b + \lambda w^T w = +1$$

$$\Rightarrow -1 + \lambda w^T w = +1$$

$$\Rightarrow \lambda = 2/w^T w$$

←

9/25/1437

Yanjun Qi / UVA CS 4501-01-6501-07

## Putting it together

- $w^T x^+ + b = +1$
- $w^T x^- + b = -1$
- $x^+ = \lambda w + x^-$
- $|x^+ - x^-| = M$
- $\lambda = 2/w^T w$

We can now define M in terms of w and b

$$M = |x^+ - x^-|$$

$$\Rightarrow M = |\lambda w| = \lambda |w| = \lambda \sqrt{w^T w}$$

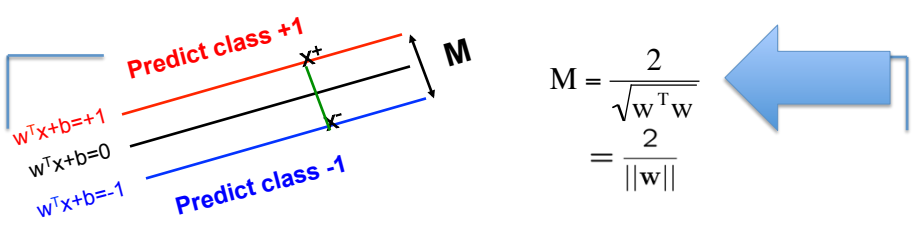
$$\Rightarrow M = 2 \frac{\sqrt{w^T w}}{w^T w} = \frac{2}{\sqrt{w^T w}}$$

←

9/25/1438

Yanjun Qi / UVA CS 4501-01-6501-07

## Finding the optimal parameters



$$M = \frac{2}{\sqrt{w^T w}} = \frac{2}{\|w\|}$$

We can now search for the optimal parameters by finding a solution that:


1. Correctly classifies all points
2. Maximizes the margin (or equivalently minimizes  $w^T w$ )

Several optimization methods can be used:  
Gradient descent, simulated annealing, EM etc.

9/25/14 39

Yanjun Qi / UVA CS 4501-01-6501-07

## Today

- Review of Classification
- Support Vector Machine (SVM)**
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  -  ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ✓ Optimization with dual form

9/25/14 40

YanJun Qi / UVA CS 4501-01-6501-07

## Optimization Step

i.e. learning optimal parameter for SVM

The diagram shows three parallel lines representing decision boundaries and the decision surface. The top line is red and labeled "Predict class +1" with equation  $w^T x + b = +1$ . The middle line is black and labeled "Decision surface" with equation  $w^T x + b = 0$ . The bottom line is blue and labeled "Predict class -1" with equation  $w^T x + b = -1$ . A point  $x^+$  is shown above the top line, and a point  $x^-$  is shown below the bottom line. A double-headed arrow between the top and bottom lines is labeled  $M$ , representing the margin. The formula  $M = \frac{2}{\sqrt{w^T w}}$  is shown to the right.

Min  $(w^T w)/2$   
 subject to the following constraints:

For all  $x$  in class + 1 }  
 $w^T x + b \geq 1$   
 For all  $x$  in class - 1 }  
 $w^T x + b \leq -1$

A total of  $n$  constraints if we have  $n$  input samples

1. Correctly classifies all points
2. Maximizes the margin (or equivalently minimizes  $w^T w$ )

9/25/14 41

## Optimization Review: Ingredients

- Objective function
- Variables
- Constraints

**Find values of the variables  
that minimize or maximize the objective function  
while satisfying the constraints**

42

Yanjun Qi / UVA CS 4501-01-6501-07

## Optimization with Quadratic programming (QP)

Quadratic programming solves optimization problems of the following form:

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

subject to n inequality constraints:

$$\begin{matrix} a_{1,1}u_1 + a_{1,2}u_2 + \dots \leq b_1 \\ \vdots \\ a_{n,1}u_1 + a_{n,2}u_2 + \dots \leq b_n \end{matrix}$$

and k equivalency constraints:

$$\begin{matrix} a_{n+1,1}u_1 + a_{n+1,2}u_2 + \dots = b_{n+1} \\ \vdots \\ a_{n+k,1}u_1 + a_{n+k,2}u_2 + \dots = b_{n+k} \end{matrix}$$

**Quadratic term**

When a problem can be specified as a QP problem we can use solvers that are better than gradient descent or simulated annealing

9/25/1443

Yanjun Qi / UVA CS 4501-01-6501-07

## SVM as a QP problem

**R as I matrix, d as zero vector, c as 0 value**

$$M = \frac{2}{\sqrt{w^T w}}$$

Min  $(w^T w)/2$

subject to the following inequality constraints:

For all  $x$  in class + 1

$$w^T x + b \geq 1$$

For all  $x$  in class - 1

$$w^T x + b \leq -1$$

A total of n constraints if we have n input samples

subject to n inequality constraints:

$$\begin{matrix} a_{1,1}u_1 + a_{1,2}u_2 + \dots \leq b_1 \\ \vdots \\ a_{n,1}u_1 + a_{n,2}u_2 + \dots \leq b_n \end{matrix}$$

and k equivalency constraints:

$$\begin{matrix} a_{n+1,1}u_1 + a_{n+1,2}u_2 + \dots = b_{n+1} \\ \vdots \\ a_{n+k,1}u_1 + a_{n+k,2}u_2 + \dots = b_{n+k} \end{matrix}$$

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

9/25/1444

## Today

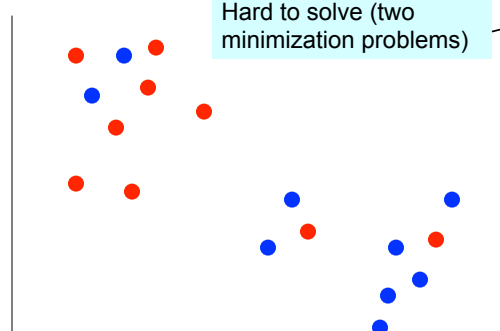
- ☐ Review of Classification
- ☐ **Support Vector Machine (SVM)**
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ➔ ✓ Non linearly separable case
  - ✓ Optimization with dual form

9/25/14

45

## Non linearly separable case

- So far we assumed that a linear plane can perfectly separate the points
- But this is not usually the case
  - noise, outliers



How can we convert this to a QP problem?

- Minimize training errors?

$$\min w^T w$$

$$\min \text{\#errors}$$

- Penalize training errors:

$$\min w^T w + C * (\text{\#errors})$$

Hard to encode in a QP problem

9/25/14

46

Yanjun Qi / UVA CS 4501-01-6501-07

## Non linearly separable case

• Instead of minimizing the number of misclassified points we can minimize the **distance** between these points and their correct plane

The new optimization problem is:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \epsilon_i$$

subject to the following inequality constraints:

For all  $x_i$  in class + 1

$$w^T x + b \geq 1 - \epsilon_i$$

For all  $x_i$  in class - 1

$$w^T x + b \leq -1 + \epsilon_i$$

Wait. Are we missing something?

9/25/14 47

Yanjun Qi / UVA CS 4501-01-6501-07

## Final optimization for non linearly separable case

The new optimization problem is:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \epsilon_i$$

subject to the following inequality constraints:

For all  $x_i$  in class + 1

$$w^T x + b \geq 1 - \epsilon_i$$

For all  $x_i$  in class - 1

$$w^T x + b \leq -1 + \epsilon_i$$

For all  $i$

$$\epsilon_i \geq 0$$

A total of  $n$  constraints

Another  $n$  constraints

9/25/14 48



Yanjun Qi / UVA CS 4501-01-6501-07

## Where we are

Two optimization problems: For the separable and non separable cases

$$\min_w \frac{w^T w}{2}$$

For all  $x$  in class + 1

$$w^T x + b \geq 1$$

For all  $x$  in class - 1

$$w^T x + b \leq -1$$

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \varepsilon_i$$

For all  $x_i$  in class + 1

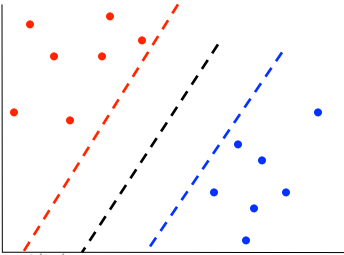
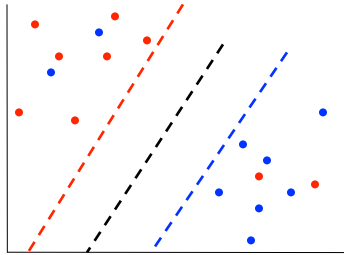
$$w^T x + b \geq 1 - \varepsilon_i$$

For all  $x_i$  in class - 1

$$w^T x + b \leq -1 + \varepsilon_i$$

For all  $i$

$$\varepsilon_i \geq 0$$

9/25/1449

Yanjun Qi / UVA CS 4501-01-6501-07

## Today

- Review of Classification
- Support Vector Machine (SVM)**
  - ✓ Large Margin Linear Classifier
  - ✓ Define Margin (M) in terms of model parameter
  - ✓ Optimization to learn model parameters (w, b)
  - ✓ Non linearly separable case
  - ➔ ✓ Optimization with dual form

9/25/1450

## Where we are

Two optimization problems: For the separable and non separable cases

$$\text{Min } (w^T w)/2$$

For all  $x$  in class + 1

$$w^T x + b \geq 1$$

For all  $x$  in class - 1

$$w^T x + b \leq -1$$

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^n C \epsilon_i$$

For all  $x_i$  in class + 1

$$w^T x + b \geq 1 - \epsilon_i$$

For all  $x_i$  in class - 1

$$w^T x + b \leq -1 + \epsilon_i$$

For all  $i$

$$\epsilon_i \geq 0$$

- Instead of solving these QPs directly we will solve a dual formulation of the SVM optimization problem
- The main reason for switching to this type of representation is that it would allow us to use a neat trick that will make our lives easier (and the run time faster)

9/25/14

51

## Optimization Review: Constrained Optimization with Lagrange

- When equal constraints
- $\rightarrow$  optimize  $f(x)$ , subject to  $g_i(x)$
- Method of Lagrange multipliers: convert to a higher-dimensional problem
- Minimize

$$f(x) + \sum \lambda_i g_i(x)$$

$$\text{w.r.t. } (x_1 \dots x_n; \lambda_1 \dots \lambda_k)$$

52

Yanjun Qi / UVA CS 4501-01-6501-07

## An alternative (dual) representation of the SVM QP

- We will start with the linearly separable case
- Instead of encoding the correct classification rule and constraint we will use LaGrange multiplies to encode it as part of the our minimization problem

Min  $(w^T w)/2$

For all  $x$  in class +1

$w^T x + b \geq 1$

For all  $x$  in class -1

$w^T x + b \leq -1$

Why? ↓↓

Min  $(w^T w)/2$

$(w^T x_i + b) y_i \geq 1$

9/25/14

53

Yanjun Qi / UVA CS 4501-01-6501-07

## An alternative (dual) representation of the SVM QP

- We will start with the linearly separable case
- Instead of encoding the correct classification rule a constraint we will use Lagrange multiplies to encode it as part of the our minimization problem

Recall that Lagrange multipliers can be applied to turn the following problem:

$\min_x x^2$

s.t.  $x \geq b$

To

$\min_x \max_\alpha x^2 - \alpha(x-b)$

s.t.  $\alpha \geq 0$

Min  $(w^T w)/2$

$(w^T x_i + b) y_i \geq 1$

9/25/14

54

## References

- Big thanks to Prof. Ziv Bar-Joseph @ CMU for allowing me to reuse some of his slides
- Prof. Andrew Moore @ CMU's slides
- Elements of Statistical Learning, by Hastie, Tibshirani and Friedman